Multivariate Data Analysis for Customer Segmentation using Principal Component Analysis and K-Means Clustering

Bosker Sinaga[1]*, Ade May Lucky Harefa[2]
Universitas Mahkota Tricom Unggul
Corresponding Author: boskersinaga@gmail.com

A R T I C L E I N F O

Keywords: Multivariate Data Analysis, Customer Segmentation, Principal Component Analysis, K-Means Clustering, Dimensionality Reduction.

A B S T R A K

This study aims to perform customer segmentation using Principal Component Analysis (PCA) and K-Means Clustering on a Mall Customers dataset consisting of 200 customers with attributes such as age, annual income, and spending score. The main problem addressed is the lack of understanding of diverse customer characteristics, leading to less effective marketing strategies. The proposed solution involves applying PCA to reduce data dimensionality and identify key features, followed by K-Means clustering to group customers into homogeneous clusters. The analysis results reveal three optimal clusters with distinct profiles: older customers with medium income and low spending scores, middle-aged customers with high income and moderate spending scores, and younger customers with low income but high spending scores. PCA visualization confirms clear segmentation, although some overlap exists near cluster boundaries. In conclusion, this segmentation provides valuable insights for companies to develop more targeted and effective marketing strategies based on the unique profiles of each customer cluster.

INTRODUCTION

The development of information technology and advances in data processing have opened up new opportunities for companies to understand customer behavior more deeply (Listy & Ilham, 2025). Data generated from customer transactions, interactions, and activities can now be collected and analyzed to generate strategic insights. One widely used form of analysis is customer segmentation, the process of grouping customers into groups with similar characteristics (Awalina & Rahayu, 2023). This

segmentation serves as a crucial basis for companies to design more targeted marketing strategies, increase customer satisfaction, and ultimately maximize profits.

Customer segmentation plays a strategic role in the modern business world because it allows companies to tailor products, services, and marketing strategies based on the specific needs of each customer group (Santoso et al., 2024). Companies that are able to identify high-value customer groups can allocate resources more efficiently and minimize marketing budget waste. Furthermore, appropriate segmentation also helps build long-term relationships with customers through personalized services and relevant communications (Ramadian et al., 2025).

However, customer segmentation is not simple, especially if the data is multivariate, meaning it consists of many interacting variables. Multivariate data often has a high dimensionality, making it difficult to analyze and visualize directly (Sihombing, 2022). The presence of correlated features can lead to information redundancy, slowing the analysis process, and reducing the accuracy of segmentation results. Therefore, an effective dimensionality reduction method is needed to simplify the data without losing important information.

Principal Component Analysis (PCA) is a statistical method widely used for dimensionality reduction in multivariate data (Shalih et al., 2025). PCA works by transforming the original variables into a new set of independent variables (principal components), which are sorted based on their contribution to the overall data variance (Bharadiya, 2023). Using PCA, data complexity can be reduced, facilitating analysis and visualization. Furthermore, PCA helps reduce noise and address the problem of multicollinearity between variables, which often occurs in customer data (Santosa, 2023).

After data dimensionality has been reduced using PCA, the segmentation process can be performed using the K-Means Clustering method. K-Means is an effective unsupervised learning algorithm for grouping data into k clusters based on the closest distance to the cluster centroid (Adawiyah & Defit, 2024). This method is known for its simplicity, computational efficiency, and ability to produce clear segmentations (Dubey & Rajavat, 2023). The combination of PCA and K-Means has been shown to improve the quality of segmentation results by minimizing noise from irrelevant variables while simplifying the visualization of results.

Although PCA and K-Means have been widely used separately, their integrated application in customer segmentation still requires in-depth study, particularly on customer data with varying demographic and behavioral variables. Many previous studies have focused solely on the use of K-Means without first optimizing the data structure through dimensionality reduction. This has the potential to result in inaccurate or difficult-to-interpret segmentations. Therefore, this study attempts to apply a combined PCA and K-Means approach to multivariate data analysis to obtain more optimal and informative customer segmentation.

This study aims to develop a customer segmentation model based on multivariate data analysis by utilizing the Principal Component Analysis (PCA) method for dimensionality reduction and K-Means Clustering for grouping. Specifically, this study aims to identify the principal components that represent the greatest variability in

customer data, group customers based on similar characteristics in the PCA space, and produce visualizations that facilitate the interpretation of each segment's profile. The results of this study are expected to contribute to the development of more effective and data-driven marketing strategies.

LITERATURE REVIEW

1. Principal Component Analysis (PCA)

   Principal Component Analysis (PCA) is a dimensionality reduction method used to transform a number of correlated variables into a number of new, uncorrelated variables, called principal components (Jolliffe & Cadima, 2016). PCA works by finding a linear combination of the original variables that can explain the greatest variance in the data. In segmentation analysis, PCA is used to simplify data without losing too much important information, as well as to simplify visualization and reduce noise (Shlens, 2014).

2. K-Means Clustering

   K-Means is an unsupervised learning algorithm used to group data into k clusters based on similarities between data points (MacQueen, 1967). This algorithm works by determining the initial centroid position, then iteratively assigning each data point to the nearest centroid and updating the centroid position until convergence (Jain, 2010). K-Means is often used in customer segmentation due to its ability to handle large data sets with relatively short computation time (Xu & Wunsch, 2005).

3. Combination of PCA and K-Means in Customer Segmentation

   The use of PCA before K-Means has been shown to improve clustering performance because PCA eliminates redundancy and different scales in variables, so that the distance between data points is more representative (Fodor, 2002). Previous research has shown that the combination of PCA and K-Means can produce clearer and more discrete segmentation than using K-Means directly on raw data (Jain & Dubes, 1988). In addition, visualizing clustering results in the principal components space facilitates interpretation and strategic marketing decision-making (Han, Kamber, & Pei, 2012).

4. Previous Studies Related to Customer Segmentation

   Several studies have applied PCA and K-Means to customer data. For example, Al-Baity and Albahli (2020) used this combination on retail data to identify high-value customer segments, while Purwanto et al. (2021) used it on e-commerce transaction data to uncover purchasing patterns. The results of these studies show that this approach is capable of identifying significantly different segments based on demographic attributes and shopping behavior.

METHODOLOGY

This research uses a quantitative data mining-based approach, combining the Principal Component Analysis (PCA) method for dimensionality reduction and K-Means Clustering for customer data grouping. The research process was conducted systematically through several main stages, starting from data collection and preprocessing, algorithm implementation, and finally, model visualization and storage.

The research was designed to segment customers based on three main variables: Age, Annual Income (k$), and Spending Score (1-100). The PCA method was used to reduce the dimensionality of the data so that highly correlated variables could be condensed into principal components without losing important information. Next, the K-Means Clustering method was used to form customer groups based on data projections into the PCA space.

The data used comes from the public dataset Mall_Customers.csv, downloaded from the Kaggle platform. This dataset contains 200 entries of customer demographic and shopping behavior information. This dataset was selected based on the completeness of relevant attributes for customer segmentation analysis. The preprocessing stage began with the selection of three relevant features from the dataset: Age, Annual Income (k$), and Spending Score (1-100). Next, a correlation analysis between the features was performed using a heatmap to examine the relationships between the variables. Afterward, feature scaling was performed using Scikit-Learn's StandardScaler to ensure each variable had a distribution with a mean of 0 and a standard deviation of 1. This process is crucial to prevent bias due to differences in scale between variables in PCA and K-Means.

The PCA method was applied to transform the original variables into uncorrelated principal components. Initially, a fit transform was performed on all features to calculate the explained variance ratio of each component, which was visualized as a scree plot. Based on the analysis results, two principal components (n_components=2) were determined to represent the data. The total variance that can be explained by these two components is calculated to ensure that missing information is minimal.

The general equation of PCA can be written as follows (Badri & Sari, 2021):

$$Z = XW \quad (1)$$

with:

Z = data matrix in principal component space
X = normalized data matrix
W = eigenvector matrix resulting from covariance decomposition

After the data is reduced to two dimensions, the K-Means Clustering method is applied with the optimal number of clusters $k = 3$. The K-Means algorithm works by minimizing the objective function (Borlea et al., 2022):

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n_j} |x_i^{(j)} - c_j|^2 \quad (2)$$

with:

$x_i^{(j)}$ = i-th data in cluster j
c_j = centroid of cluster j

k = number of clusters
n_j = number of members in cluster j

Visualization of the clustering results is carried out in three forms, namely scatter plots in PCA space with coloring based on clusters and marking of centroid points to see the separation of groups clearly, pair plots on the original features used to observe the distribution of variables in each cluster, and bar plots that display the average of each feature in each cluster as a customer segmentation profile.

RESEARCH RESULT

This study used 200 customer data with five main attributes: CustomerID, Gender, Age, Annual Income (k$), and Spending Score (1–100). The entire analysis process was carried out in stages: data exploration, standardization, PCA dimension reduction, determining the number of clusters, clustering using K-Means, and cluster profile analysis.

3.1. Initial Data Description

The dataset used is the Mall Customers Dataset from Kaggle, containing 200 shopping mall customer data. The dataset includes five variables: CustomerID (customer's unique identifier), Gender (gender), Age (age in years), Annual Income (annual income in thousands of US dollars), and Spending Score (shopping behavior score from 1–100).

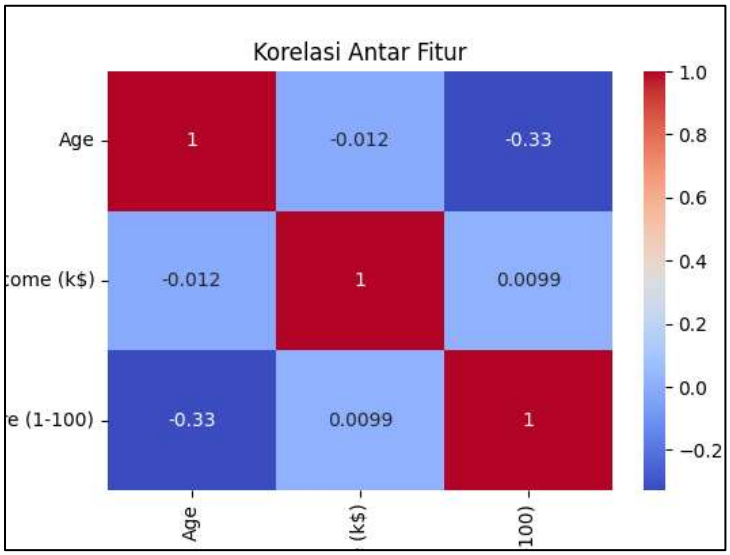


Figure 1. Intervariable Correlation Heatmap

Figure 1 shows that the correlation between the numerical variables is relatively low. The Age variable has a moderate negative correlation with the Spending Score of -0.33, indicating that the older a customer is, the lower their spending score tends to be. Meanwhile, Annual Income has a very low correlation with both Age (-0.012) and Spending Score (0.0099). This indicates that each variable contributes unique information,

making it suitable for use together in subsequent analysis without significant risk of multicollinearity.

3.2. Data Standardization and PCA Analysis

Before conducting further analysis, the numerical data for the Age, Annual Income, and Spending Score variables were first standardized using the Z-score normalization method. The purpose of this standardization is to ensure that each variable has a mean of zero and a standard deviation of one, so that differences in scale between variables do not affect the analysis results. This is important because methods such as PCA are very sensitive to differences in data scale. After the data was standardized, Principal Component Analysis (PCA) was performed to reduce the data dimensionality and identify linear combinations of variables that can explain the greatest variation in the dataset. PCA helps simplify data complexity, reduce noise, and facilitate visualization, without losing too much important information.

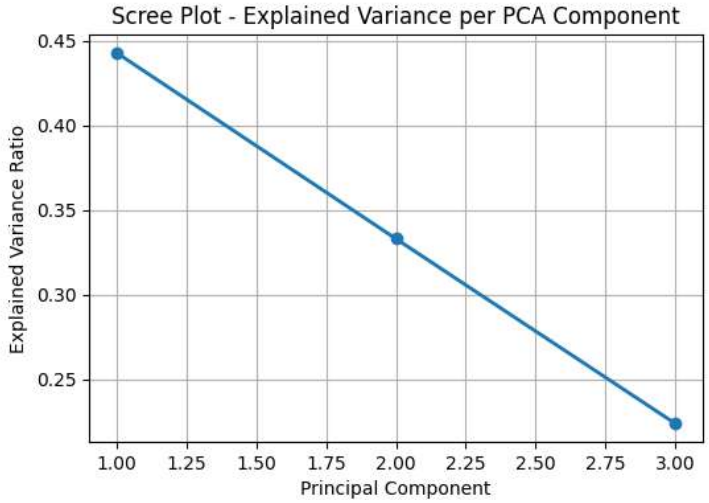


Figure 2. Scree Plot – Explained Variance per PCA Component

The PCA results are shown in Figure 2, in the form of a Scree Plot, which illustrates the explained variance ratio for each principal component. Based on the graph, Principal Component 1 (PC1) explains approximately 44.7% of the data variation, Principal Component 2 (PC2) explains 33.4%, and Principal Component 3 (PC3) explains 21.9%. Thus, the first two principal components are able to explain approximately 78.1% of the total data variation, which is representative enough to be used in the visualization and clustering analysis stages.

3.3 Visualization of Clustering Results with PCA

After determining the optimal number of clusters using the Elbow method, a clustering process using the K-Means algorithm was performed on the data, which had been reduced in dimension using PCA. This dimensionality reduction aims to simplify visualization and interpretation while ensuring that emerging patterns are based on the primary variations in the data.

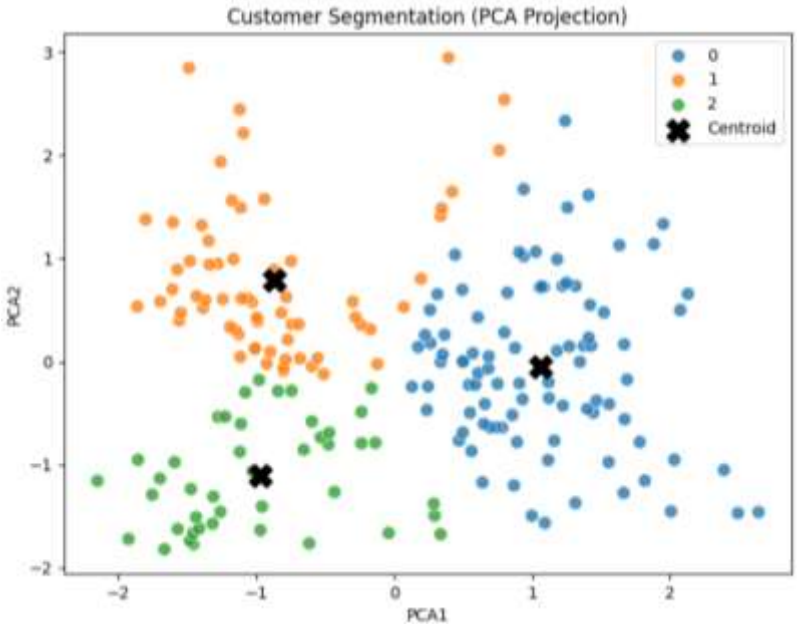


Figure 3. Visualization of K-Means Clustering in PCA Space

Figure 3 shows the results of customer segmentation in two-dimensional space based on Principal Component 1 (PCA1) and Principal Component 2 (PCA2). Each point on the graph represents one customer, with different colors indicating membership in each cluster, namely Cluster 0 (blue), Cluster 1 (orange), and Cluster 2 (green). The large black cross (X) in each cluster indicates the position of the centroid, which is the center point of each cluster calculated based on the average position of all cluster members in the PCA space. This centroid position is used by the K-Means algorithm to determine boundaries between clusters and minimize the average distance between points within the cluster. Based on the visualization, it can be seen that the distribution of each cluster is relatively separate, although there are several points that are in the boundary area and adjacent to other clusters. This indicates that the resulting segmentation is quite effective in distinguishing customer patterns based on the variables analyzed.

3.4 Visualization of Clustering Results

After determining the number of clusters, clustering was performed using the K-Means algorithm on the PCA space (PC1 and PC2). The results were visualized as a scatterplot with different coloring for each cluster and a special mark for the cluster centroid.

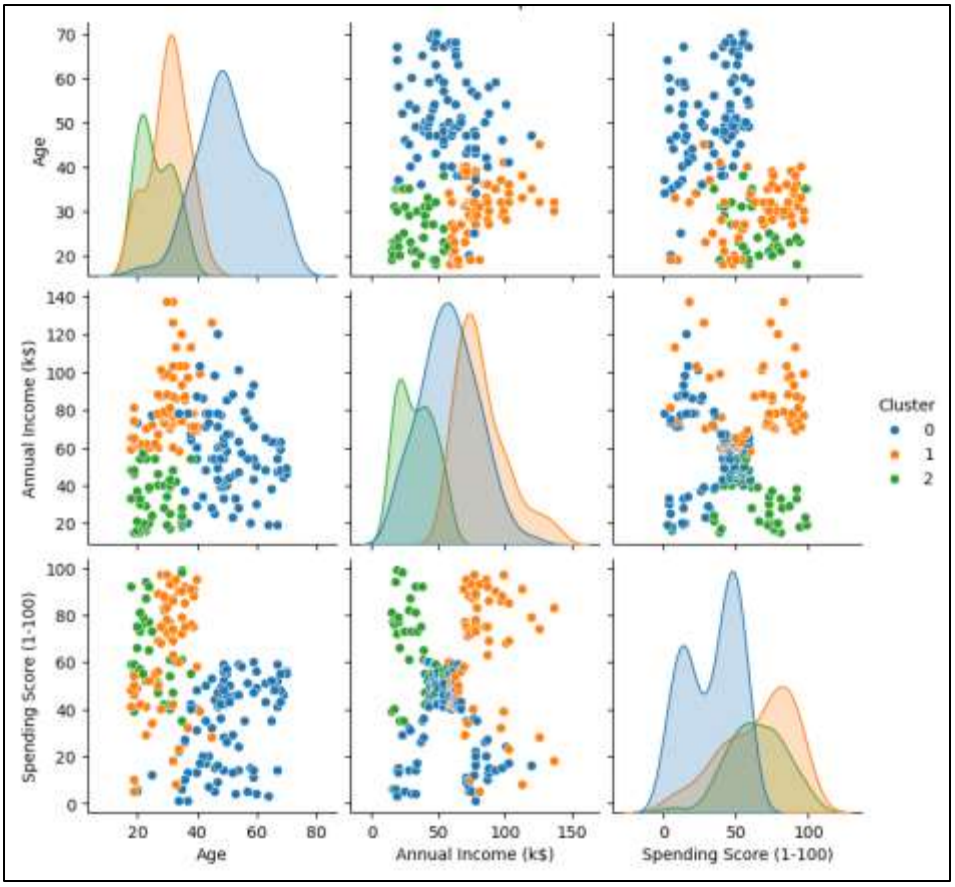


Figure 4. Visualization of Grouping Results

Based on Figure 4, it shows the relationship between Age, Annual Income (k$), and Spending Score (1–100) which have been grouped using the K-Means method into three clusters: Cluster 0 (blue), Cluster 1 (orange), and Cluster 2 (green). The diagonal graph displays the distribution of each variable, while the other graphs show the relationship between variables. In general, Cluster 0 contains older customers with varying income and spending scores, Cluster 1 is dominated by middle-aged customers with high income and moderate spending scores, while Cluster 2 contains young-middle-aged customers with low income and varying spending scores.

3.5 Customer Cluster Profile Analysis

The average profile of each customer feature formed from the K-Means clustering process provides an overview of the unique characteristics of each group. The variables analyzed include Age, Annual Income (k$), and Spending Score (1–100). The average differences in each variable reveal differences in customer behavior and potential value, which can be utilized in marketing strategy planning.

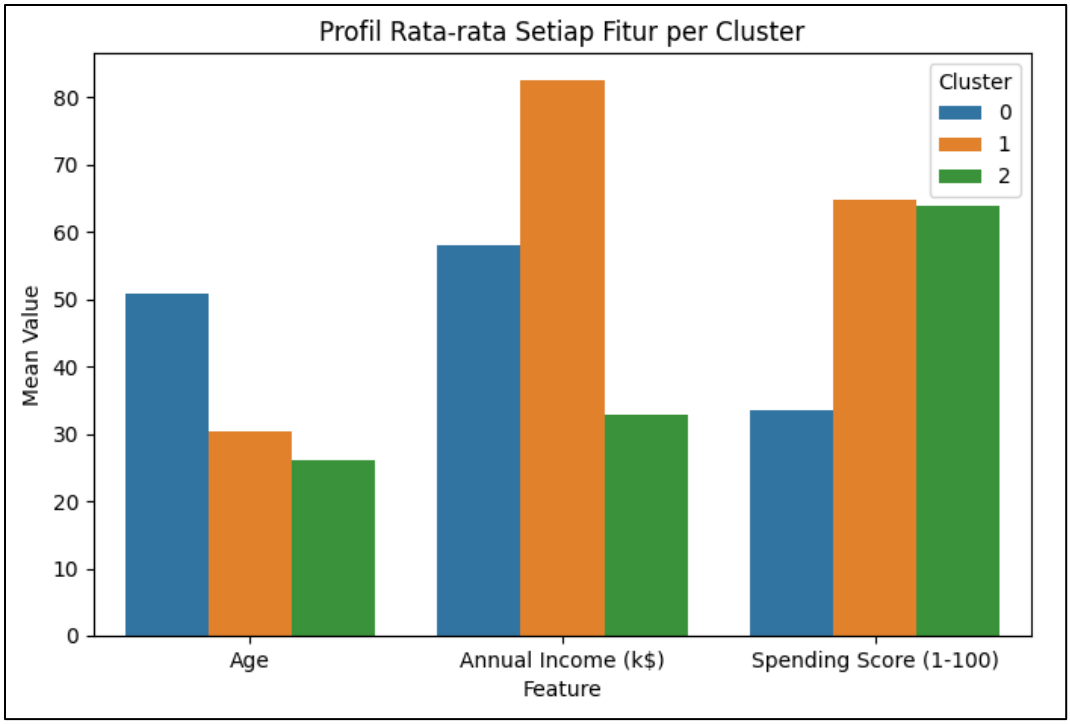


Figure 5. Average Profile of Each Feature per Cluster

The visualization in Figure 3.5 shows that Cluster 0 (blue) has the highest average age, medium annual income, and low spending scores. Cluster 1 (orange) stands out with the highest annual income, medium age, and relatively high spending scores, making it a high-value customer segment. Meanwhile, Cluster 2 (green) consists of customers with the youngest age, lowest income, but high spending scores. These differences indicate the need for different marketing approaches for each segment, so that the implemented strategy is more effective and targeted.

DISCUSSION

This study shows that the combination of PCA and K-Means is effective for customer segmentation. The two main components of PCA explained 78.1% of the data variation, simplifying visualization without losing important information. The clustering results yielded three segments with distinct characteristics:

- Cluster 0: oldest age, middle income, low spending score.
- Cluster 1: middle age, highest income, high spending score (high-value customers).
- Cluster 2: youngest age, low income, high spending score.

This segmentation provides a strategic basis for more targeted marketing, such as exclusive promotions for high-value customers, affordable pricing for younger customers, and retention programs for older customers.

CONCLUSIONS AND RECOMMENDATIONS

This study successfully segmented customers using the K-Means clustering method combined with Principal Component Analysis (PCA) dimension reduction. The three clusters formed had distinct characteristics based on age, annual income, and customer shopping behavior scores. Cluster 0 consisted of customers with the oldest age and low shopping scores, Cluster 1 consisted of customers with the highest income but moderate shopping scores, while Cluster 2 contained the youngest customers with the lowest income but high shopping scores. This segmentation demonstrated differences in customer behavior patterns that can be used as a basis for more targeted marketing strategies. Visualization of the PCA results and scatter plots confirmed a fairly clear cluster division although there was slight overlap in the boundary areas between the groups.

ADVANCED RESEARCH

Based on the research results, there are several development directions that can be taken to improve the accuracy and relevance of customer segmentation: Addition of Behavioral and Demographic Variables, Use of Alternative Clustering Methods, Integration of Predictive Analysis, Evaluation of Cluster-Based Marketing Strategy Performance, Use of Non-Linear PCA or Other Dimensionality Reduction Techniques, Dynamic Analysis Using Periodic Data.

REFERENCES

Adawiyah, Q., & Defit, S. (2024). Penerapan Algoritma K-Means Clustering untuk Mengelompokkan Rekomendasi Metode Kontrasepsi Berbasis Machine Learning di Puskesmas. *Jurnal KomtekInfo*, 300–305.

Awalina, E. F. L., & Rahayu, W. I. (2023). Optimalisasi Strategi Pemasaran dengan Segmentasi Pelanggan Menggunakan Penerapan K-Means Clustering pada Transaksi Online Retail. *Jurnal Teknologi Dan Informasi*, *13*(2), 122–137.

Badri, F., & Sari, S. U. R. (2021). Penerapan metode Principal Component Analysis (PCA) untuk identifikasi faktor-faktor yang mempengaruhi sikap mahasiswa memilih melanjutkan studi ke Kota Malang. *Build. Informatics, Technol. Sci*, *3*(3), 426–431.

Bharadiya, J. P. (2023). A tutorial on principal component analysis for dimensionality reduction in machine learning. *International journal of innovative science and research technology*, *8*(5), 2028–2032.

Borlea, I.-D., Precup, R.-E., & Borlea, A.-B. (2022). Improvement of K-means cluster quality by post processing resulted clusters. *Procedia Computer Science*, *199*, 63–70.

Dubey, P., & Rajavat, A. (2023). Effective K-means clustering algorithm for efficient data mining. *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, 1–6.

Listy, V., & Ilham, I. (2025). Revolusi Sistem Informasi Manajemen di Era AI dan Big Data Mengubah Cara Bisnis Bekerja. *Simpatik: Jurnal Sistem Informasi dan Informatika*, *5*(1),

27–36.

Ramadian, A., Judijanto, L., & Erwin, E. (2025). *Customer Relationship Management (CRM): Strategi Membangun Hubungan Pelanggan yang Kuat*. PT. Green Pustaka Indonesia.

Santosa, Y. P. (2023). Kombinasi Linier Target Data Untuk Regresi Multitarget Menggunakan Principal Component Analysis. *Jurnal Teknologi Terpadu*, *9*(1), 1–9.

Santoso, R. P., Ningsih, L. S. R., & Irawati, W. (2024). Implementation Of Segmenting Targeting And Positioning Strategies In Improving Marketing Performance. *BIMA: Journal of Business and Innovation Management*, *6*(2), 280–292.

Shalih, F. A., Ramadhan, R. A., & Syalaisa, N. (2025). Comprehensive Overview of Principal Component Analysis Applications and Developments. *Jurnal EurekaMatika*, *13*(1), 25–34.

Sihombing, S. O. (2022). *Pengantar metode analisis multivariat*. Penerbit NEM.